DOCUMENT RESUME

ED 364 578 TM 020 802

AUTHOR Burton, Elizabeth; Linn, Robert L.
TITLE Report on Linking Study--Comparability across Assessments: Lessons from the Use of Moderation Procedures in England. Project 2.4: Quantitative Models To Monitor Status and Progress of Learning and Performance.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Aug 93
CONTRACT R117G10027
NOTE 39p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; *Achievement Tests; *Comparative Analysis; *Educational Assessment; Elementary Secondary Education; Foreign Countries; National Programs; *Student Evaluation; Systems Development; Test Construction; Testing Programs; *Test Results
IDENTIFIERS England; Linking Exam Results; *Linking Metrics; *National Information Systems; United States

ABSTRACT

Currently in the United States there is considerable interest in developing a system of examinations that is substantially different from the large-scale testing programs presently in use in this country. These new exams would involve a different form of assessment than current tests and would also enable greater local control over the assessments. Results of such exams would then need to be linked in some way to national standards in order to permit comparisons across schools and regions. This paper discusses some of the issues and problems involved in linking the results of such exams; studies and papers that describe experiences with linking results from the different examination boards in England provide the basis for the discussion. The two major approaches to linking exam results that have been used in England involve either the use of external exams and statistical linking methods, or judgmental audits. Advantages and problems of each of these approaches, and reasons that neither approach is satisfactory by itself, are discussed. Instead, some combination of the two would seem to be necessary. Such a system may involve use of external exams and statistical procedures to identify places where results of the exam used locally may be out of line with national standards. These cases could then be resolved by gathering additional information to determine whether the discrepancies were valid, or by performing some type of judgmental audit. (Contains 20 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 364 578

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – August 1993

Project 2.4 Quantitative Models to Monitor
Status and Progress of Learning and Performance

Report on Linking Study – Comparability Across Assessments: Lessons From the Use of Moderation Procedures in England

Robert L. Linn, Project Director
CRESST/University of Colorado at Boulder

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

TM020802

The work reported herein was supported in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

BEST COPY AVAILABLE

COMPARABILITY ACROSS ASSESSMENTS: LESSONS FROM THE USE OF MODERATION PROCEDURES IN ENGLAND

Elizabeth Burton and Robert L. Linn
CRESST/University of Colorado at Boulder

Abstract

Currently in the United States, there is considerable interest in developing a system of examinations that is substantially different from the large-scale testing programs currently in use in this country. These new exams would involve a different form of assessment than current tests and would also enable greater local control over the assessments used. The results of such exams would then need to be linked in some way to national standards in order to permit comparisons across schools and regions. This paper will discuss some of the issues and problems involved in linking the results of such exams; studies and papers that describe experiences with linking results from the different examination boards in England provide the basis for the discussion.

The two major approaches to linking exam results that have been used in England involve either the use of external exams and statistical linking methods, or judgmental audits. The advantages and problems of each of these basic approaches, and the reasons that neither approach is satisfactory by itself, are discussed. Instead, some combination of the two approaches would seem to be necessary. Such a system may involve the use of external exams and statistical procedures to identify places where the results of the exam used locally may be out of line with national standards. These cases could then be resolved by gathering additional information to determine whether the discrepancies were valid, or by performing some type of judgmental audit.

In the last few years, considerable interest has been expressed in the development of a system of assessments that would be radically different from any of the large-scale testing programs that flourished during the past several decades in the United States. Features of the system are illustrated by proposals advanced by the National Education Goals Panel (1991), the National Council on Education, Standards and Testing (1992), the New Standards Project (Learning Research and Development Center & National Center for Education

and the Economy, 1992), and in plans for new state assessments, as well as by some of the features of pending legislation (Goals 2000: Educate America Act).

In brief, the type of system that is envisioned differs from testing programs of the past in two major ways. First, the form of assessment would undergo a major change. Second, the system would allow for local or state variations in the specific assessments that would be used, but the assessment results would all be linked together in some, as yet, not explicitly identified way so that student and system performance could be compared to common national performance standards.

The focus of this paper is on the second of these new features, the linking to common standards. Because approaches to dealing with the problem of comparing local or state assessments to national standards are affected by the properties of the assessments, however, it is worth highlighting some of the key characteristics of proposed changes in the form of assessment.

There is a rapidly growing movement away from an almost exclusive reliance on multiple-choice tests toward assessments that require extended student responses, performances, projects, or collections of student work. The rationale for this shift in the form of assessment has been discussed by a number of authors (e.g., Linn, Baker, & Dunbar, 1991; Resnick & Resnick, 1992, Shepard, 1991; Wiggins, 1989) and need not be repeated here. Five interrelated characteristics of these alternative forms of assessment are relevant for the linking problem, however, and therefore deserve mention. Among other things, the assessments being proposed generally have the following characteristics: (a) Each task requires a relatively extended period of time to administer; (b) because of the length of time required per task, relatively few tasks can be administered to any given student; (c) there is not one "right" or best answer, process, or product; (d) because tasks can be approached in many ways and a variety of responses may be considered excellent, judgmental scoring of student responses or products is required; and (e) in many instances, students or teachers may choose the tasks to be performed.

The above characteristics have implications for the goal of comparing results of different assessments to common national standards. In particular, statistical approaches that have been used to equate alternate forms of standardized tests are not applicable to the problem posed by the proposed

assessment system (see, for example, Linn, 1993; Mislevy, 1992). Some other approach to linking will be needed if the goal of making comparisons to a common national standard is to be realized.

The linking problem of proposals for a system of assessments is similar in some respects to the problem of achieving comparability across different examining boards in England and some other countries. The general approach that has been used in England to achieve an acceptable degree of comparability is referred to as "moderation." The basic idea is that a procedure is established to identify instances of local scoring that is overly stringent or overly lenient in comparison to other boards and, once identified, to "moderate" those scores to bring them more into line.

There are a number of specific approaches to moderation, some of which rely on external exams and statistical moderation and others that depend on judgmental audits or consensus moderation procedures. Although it is unlikely that a moderation procedure used in England would fit in every detail the context of proposals for a system of assessments tied to a common national standard in the United States, it seems likely that there is much that can be learned about the strengths and weaknesses of various moderation procedures used in England. It is in this spirit that the present review of moderation procedures and investigations of those procedures was undertaken.

Background

In the mid-1960s the Schools Council in England first began planning to implement a system of internal assessment in secondary schools. An important part of internal assessments has been the regional and national moderation of the results of those assessments. Internal, instead of external, assessments were seen to be advantageous for two reasons. First, teachers' assessments would be based on a long-term knowledge of each student and his/her work and, therefore, it was believed that they would be more valid than an assessment made on the basis of a particular exam given on a particular day. Second, local control over syllabuses would be enhanced since teachers wouldn't have to worry about teaching to an external test that might not measure the abilities and skills that they or their school or district considered to be most important. However, in order to enable comparisons of individual or aggregated results across schools

and regions and to ensure public belief in the validity of internal assessments and acceptance of their results, some type of moderation of scores is necessary.

What follows are summaries of papers or sections of papers that discuss experiences with and studies of various approaches to moderation. The summaries fall into three sections: The first includes procedures that can be classified as moderation by inspection or cross-moderation (in which judgmental audits are used); the second, procedures based on statistical moderation (in which a reference or monitor examination or other external information is used); and the third includes procedures that are enhancements to one of these general methods or a combination of the two. The information comes not only from studies on moderation in secondary school examinations, but from vocational and business and technical examinations as well.

Currently in England, secondary school exams in various subjects are developed and administered by nine examination boards. Individual schools are free to choose the examination board that fits best with their priorities and standards. This situation allows local control to be maintained while providing a way to verify that the examinations used are of high quality. A major problem with this system is that a means of ensuring the comparability of scores across boards is needed. The various studies discussed below describe attempts to provide such assurance through different moderation procedures. The purpose of this paper is to describe the various procedures, to draw attention to their complexities, and to highlight the trade-offs that any given procedure entails. Implications of the findings for systems that are under consideration in the U.S. will also be discussed.

Moderation by Inspection

Teacher-moderator comparisons. Moderation by inspection procedures are based primarily on moderators' inspection (review, re-grading, or independent grading) of the students' "scripts" (where a script refers to the total collection of work on which the student is graded and can include exam results, papers, projects, and performances) and ratification or repudiation of the awarded grade. Early studies of moderation by inspection describe a process whereby moderators re-grade a sample of scripts that have already been graded by the teacher. Cohen and Deale (1977) recommend that the scripts be chosen either randomly or by the moderator rather than the teacher so that the teachers

are not able to obtain higher grades for their students by grading the sample severely. While the specifics of different moderation by inspection procedures differ somewhat, many aspects are common to most or all of them. In most cases, trial marking sessions are held and are attended by teachers (Smith, 1978, p. 9), moderators (Schools Council, 1965, p. 3), or both (Cohen & Deale, 1977, pp. 46-47). As noted by Smith (1978), the results of these sessions, along with decisions about criteria, need to be carefully documented. One aspect of these early moderation procedures that appears to have been abandoned in more recent applications is that all of a teacher's grades were adjusted up or down if the moderator(s) found them to be too severe or lenient (Schools Council, 1965, p. 12). This process not only gives greater weight to the moderator's judgment than to the teacher's, but also results in problems if the moderator believes that the teacher's grades are invalid (i.e., if the moderator disagrees with the rank order established by the teacher's grades). In circumstances where it is not possible to re-grade all of the students' scripts, a choice has to be made between excluding the teacher-graded component of the assessment[1] and ignoring the discrepancy in rankings and adjusting for overall leniency or severity of the teachers' grades only (Cohen & Deale, 1977, p. 48).

Approaches to verifying the judgment of the moderators range widely across moderation procedures. In some cases, there doesn't seem to be any formalized procedure although it is acknowledged that it is "important that some sort of random check is carried out on moderators' standards even when, on the surface, all appears to be satisfactory" (Cohen & Deale, 1977, p. 45). At the opposite extreme are procedures in which the moderators' judgments are verified by a Review Panel in all cases, regardless of whether the moderator and teacher agree (Smith, 1978, p. 10). Interestingly, moderation procedures that give the most power for decision making to the moderators—for the most part, early moderation attempts—appear generally to have involved the least amount of verification of their judgments.

---

[1] Smith (1978) and Cohen and Deale (1977) describe moderation procedures that are used in the 16+ examinations, in which the teacher-assessed component is only one component of the total exam grade. The teacher-assessed component comprises a different percentage of the total exam for different subjects and examination boards. Ignoring the teacher-assessed component of the exam and basing the grade on the external component alone will be more or less feasible depending on the weight given each of the two components.

Applications of moderation by inspection vary substantially with regard to the specificity of the criteria used for comparing the assessments of teachers and moderators. In some cases, it appears that the decision as to whether the teachers' grades were out of line was left largely to the moderator without specific criteria being delineated (Cohen & Deale, 1977, p. 47). In other cases, very specific criteria are recommended. One suggested procedure, described in detail by Smith (1978, pp. 15-18), involves using the ranges of scores given by teachers and moderators to set tolerance limits on three characteristics; if any one of the three criteria is not met, the scripts are referred to a second moderator to determine whether grade adjustments should be recommended. The three characteristics are: (a) *discrimination*—the ranges of scores awarded by the teacher and moderator must be sufficiently similar (i.e., within the selected tolerance limits); (b) *standard*—the actual scores must be sufficiently similar (i.e., neither scorer should award grades that are consistently higher or lower than those given by the other scorer); and (c) *conformity*—the rank orders of the students established by the two scorers should be similar (Smith, 1978, p. 15).

Clearly the tolerance limits selected when using this procedure are arbitrary to some extent; they "represent a compromise between the ideal of perfect agreement (for which the critical values would be zero) which is unattainable and values which would be so large that the number of occasions on which the requirements would not be met would be few." The justification for using these particular values is that "in practice [they] have been found to be acceptable" (p. 18). Those wishing more information about the rationale and procedures for the use of range estimates are referred to the Schools Council's Examinations Bulletin No. 5 (Schools Council, 1965. *The Certificate of Secondary Education: School-based examinations* [Examinations Bulletin No. 5]. London, HMSO).

These grade comparisons serve two purposes: One is to determine which teachers are sufficiently in line with the consensus view to be considered for the role of moderator in a subsequent assessment; the other is to identify teachers whose grades deviate substantially from those assigned by the moderator (p. 16). If a teacher's grades are found to be out of line according to the criteria delineated above, the scripts are referred to a second moderator or Chief Moderator, who decides whether all of the teacher's grades should be adjusted (p. 18).

Some additional issues and questions, many of them unresolved, are mentioned in these early studies. One recommendation that is made is to provide opportunities for teachers to discuss specific scripts and grading criteria with moderators, especially in cases where the moderators' judgments are used to make adjustment decisions (Cohen & Deale, 1977, pp. 46-47; Smith, 1978, p. 9). One question that arises is if the scripts should be sent to the moderator or if the moderator should visit the school. In some cases, the nature of the assessment makes it necessary for the moderator to visit the school (for example, if the assessment involves a performance or a "continuous assessment"). However, it is recognized that even with several visits to the school, the moderator's assessment may be based on too light a sample of the students' work to be valid (Cohen & Deale, 1977, p. 47; Smith, 1978, pp. 8-9). In addition, in many cases, the moderation process may involve far too many students, schools, subjects, etc. to make visits to the schools feasible (as appears to be the case in most recent cross-moderation studies). A final question that has not yet been explored is what the effect is on the moderators' grades if they know the grades given by the teachers or if the teachers' comments remain on the scripts (Cohen & Deale, 1977, p. 45; Smith, 1978, pp. 14-15).

Smith discusses the results of a survey given in 1975 to teachers who were involved in an assessment scheme that used a moderation by inspection procedure. Teachers acknowledged that the duties involved in assessment and moderation were very difficult to fit into an already busy schedule but, nevertheless, "there was enthusiastic support not only for the [assessment] scheme but also for the method of moderation." In general, they preferred moderation by inspection to statistical moderation (more below) (Smith, 1978, pp. 18-19).

Cross-moderation. The remaining studies in this section discuss the process of cross-moderation, which involves the use of moderation by inspection procedures but focuses specifically on ensuring the comparability of exam grades given by the different examination boards. In this procedure, examiners from each board moderate scripts from the other boards involved in the study. (One could imagine an analogous application in the U.S. being cross-state moderation in response to the push for referencing state assessment results to national performance standards.) A number of problems that remained implicit in earlier

studies are discussed more explicitly in these studies. Bardell, Forrest, and Shoesmith (1978) point out that

> [c]ross-moderation studies are founded on the assumption that subject experts . . . , on the basis of their professional judgments and despite the numerous differences between examinations, can decide from a scrutiny of scripts whether comparable grades are being awarded by the boards to candidates of comparable levels of attainment. (p. 27)

They discuss as well the tension that exists between dealing with a reasonable number of syllabuses and exams and having a full national context on which to base judgments (p. 27). Johnson & Cohen (1983) point out that the method of cross-moderation "has come in for much criticism for lack of rigour and, in particular, for failure to produce conclusive results" (p. vii). One major problem with trying to ensure comparability is defining what comparability *is*: Are the scores meant to be comparable as measures of general capability or of attainment of specific goals in a subject? This distinction, clearly, is the familiar one between aptitude and achievement definitions of assessments (p. 1). Another major problem identified is the difficulty moderators have in agreeing upon standards and stating them explicitly (Bardell et al., 1978, p. 29). For example, given that the major purpose of cross-moderation is to ensure comparability between the examinations of different boards, should the standards and grading criteria used in making judgments be those of the moderator, the moderator's board, or the board from which the scripts are taken (Bardell et al., 1978, p. 29; Johnson & Cohen, 1983, pp. 16-17)? And difficulties in establishing comparability arise not just from different grading standards, but from differences in syllabuses, modes of assessment, marking schemes, and relative weightings given to the components (Bardell et al., 1978, p. 15; Forrest & Shoesmith, 1985, pp. 26-29; Johnson & Cohen, 1983, pp. 5-6, 11-16; Schools Council, 1965, p. 15). In addition, students from different boards may have different levels of ability (Johnson & Cohen, 1983, p. 3). Syllabuses may emphasize practical or theoretical knowledge. Examinations may use different modes of assessment (written, oral, performance, etc.), may require different numbers of papers, and may include optional papers or a choice of questions. Different abilities and skills may be emphasized by different boards. Finally, some exam evidence may be missing at the time of moderation (for example, if oral or performance components are included). The problem of missing evidence

is compounded by the fact that most students do not perform consistently across tasks. While it is possible to choose students for the study who *do* perform consistently across tasks, these students are not likely to be a representative group and the validity of the study will be lessened (Johnson & Cohen, 1983, pp. 11-16).

These differences in the boards' exams make comparisons and quantification of differences in standards extremely problematic. Forrest and Shoesmith (1985) identify a tension that arises between increasing comparability on the one hand and maintaining local control over the skills and abilities emphasized in the tests on the other. One potential way of dealing with this problem is to introduce both general and subject-specific criteria into the assessment guidelines (p. 26).

Although there are quite a few problems with trying to establish comparability, Bardell et al. (1978) point out that cross-moderation procedures also have a number of advantages (pp. 30-31). First, cross-moderation is most like the actual task a board faces when awarding grades in that it relies on the scripts themselves rather than outside tests and it utilizes the examiners who are responsible for syllabuses and exam papers. On the other hand, Forrest and Shoesmith (1985) point out, cross-moderation requires separation of skills (e.g., explicitly addressing the students' abilities with regard to each of a number of specified criteria and combining these assessments into an overall grade) while, operationally, assessors tend to grade holistically (p. 28). The second advantage of cross-moderation procedures is that they allow examiners from different boards to get together and discuss grading standards and other pertinent issues. Third, the exercises themselves often help to reduce differences among the participants in the standards they apply. The main disadvantages of cross-moderation, in addition to establishing comparability, are that it is very time-consuming and vulnerable to a number of reliability and validity problems (Bardell et al., 1978, p. 30). While it is not reasonable to require assessors to grade a large number of scripts, including too few scripts will increase the sampling error and may give inaccurate results. In addition, assessors are dealing with an unfamiliar exam which may lead to reliability problems. Finally, the judgments of the assessors are necessarily quite subjective. A related point made by Forrest and Shoesmith (1985) is that time pressures may make it tempting not to include scripts from every board in the study; if this is

done, of course, the results are not generalizable beyond those boards that are included (pp. 29-30).

Cross-moderation studies are categorized into two general types: (a) identification studies, which involve sampling a wide range of ungraded scripts and identifying grading criteria and the locations of grade borderlines; and (b) ratification studies, in which representative samples of scripts that have teacher-awarded grades at specified borderlines are examined in order to verify that borderlines are appropriately defined and accurately identified (Bardell et al., 1978, p. 28; Forrest & Shoesmith, 1985, p. 25; Johnson & Cohen, 1983, p. 5). Johnson and Cohen (1983) point out that because identification studies require the examination of a large number of scripts, and, consequently, a large time commitment, a method of ratification in which batches of scripts from different boards are rank-ordered to determine which boards' grades tend to be too lenient or severe has become increasingly popular (p. 5). However, the authors continue, gaining an advantage in terms of the reasonableness of the task entails some trade-offs (pp. 21-23). First, ranking of scripts does not offer any procedure for quantifying differences in boards' standards; past attempts to quantify differences using a 5-point leniency to severity scale have failed to produce unambiguous results. Second, the reliability of grading judgments depends on the representativeness of the set of scripts used; because of the relatively small number of scripts used in ratification studies, representativeness is likely to be a problem. A related point is that typical ratification studies do not allow quantification of the degree of confidence that can be placed in their results, since the distributions of original grades in the samples used are not consistent across boards. Finally, the results of the exercise could be different if either a different set of scripts were used (Forrest & Shoesmith, 1985, p. 29; Johnson & Cohen, 1983, p. 23) or a different set of assessors were involved (Johnson & Cohen, 1983, p. 23). Johnson and Cohen's study was designed to explore potential solutions for some of these problems.

One of the questions that Johnson and Cohen's (1983) study addressed was that of what grading criteria should be used, those of the individual moderators, the moderators' boards, or the board from which the scripts are taken (p. 16). The authors argued that grade criteria are too complex to expect examiners to consistently represent their board's criteria across all grade levels (pp. 17-18). They suggested, therefore, that examiners keep in mind the aspects of

performance valued by the source board, which can be gathered from the boards' exam materials (p. 19). These performance qualities are relatively easy to identify. While grading criteria are much more difficult to articulate (for example, deciding what *balance* of different qualities is appropriate for a given grade), examiners need only to apply their implicit grading criteria consistently throughout the exercise (pp. 20-21). It is possible that assessors would get a better knowledge of the grade scheme they were to use if they first marked the scripts using the appropriate marking scheme, which explicitly lists the performance criteria to look for (p. 25). In order to explore this possibility, initial immediate-impression grades were to be compared to grades given after marking (p. 26).

Another problem that Johnson and Cohen addressed was that of quantifying differences in boards' standards (pp. 24-26). The usual procedure in ratification studies is to compare average assigned scores to average original scores; however, this procedure offers no way to quantify the differences if the original score spread varies across the batches of scripts from different boards. In order to quantify and assess reliability, distributions of original scores must be similar, and scripts should be spread evenly across score ranges to see if differences in grading standards are consistent across all scores. The consistency of differences across score levels would also be assessed by examining batch differences separately at each level. Finally, using more than one assessor from each board would allow a determination of the extent to which being a member of a particular board affects scoring. A board effect could be inferred if all assessors from a given board gave scores that were consistent with each other's but inconsistent with scores given by assessors from other boards. The identification of any of the above effects was limited in the Johnson and Cohen study by at least two design constraints (p. 27): (a) the limited availability of examiners who had the necessary knowledge and experience to take part in grading exercises of this kind; and (b) the workload required of each assessor.

The study was conducted as follows: Each board received two matched samples of scripts, one for immediate-impression and the other for grading-after-marking grades. First, they reviewed all material except the marking schemes, then did the immediate-impression grading; the grading-after-marking was done later by each assessor at his or her home. Assessors met after the residential grading to discuss differences in syllabuses, whether the process of grading led

them to change their expectations for the scripts, and the criteria they used in grading. In addition, assessors were invited to write in with any comments or problems they had in grading-after-marking. The study was done for a different level of achievement in each of three subjects: physics, French, and mathematics. Assessors from three boards participated in the study, using scripts obtained from the same three boards.

The differences in the standards of the three boards that participated were assessed by averaging the grades given by all three groups of assessors to each batch of scripts (so that an average grade for each board's scripts was computed) and comparing those three averages. The consistency of these differences across score levels was also investigated. Differences in grading standards of the three groups of assessors (i.e., assessors from certain boards are more or less severe than those from other boards) would indicate that it is difficult for assessors to adopt the standards used by other boards; comparing the grades of assessors within boards would indicate how consistent this board effect is. Differences between immediate-impression and grading-after-marking results would indicate that marking improved the assessors' ability to apply another board's standards. Finally, any differences in the ranking of scripts from different boards, by assessors within or between boards, would indicate that the assessors' perceptions of each board's grading standards were not reliable.

Based on their analyses, Johnson and Cohen concluded that, in general, there was no overall tendency on the part of the assessors to be lenient or severe, suggesting that assessors were able to apply the standards of the different boards. Furthermore, there were dissimilarities in the grading standards of the three assessors from any particular board. That is, differences in assessor stringency within boards were similar to those between boards. They also speculated that it is possible that the tendency to be severe that has been found in previous studies (Johnson & Cohen, 1983, p. 63; Forrest & Shoesmith, 1985, p. 33) was due to instructions to assessors to apply their own board's standards (either because assessors believe that the criteria of their board are superior to those of the others and consciously or subconsciously grade scripts from other boards more severely or because they are looking for the abilities stressed by their board which may not be stressed by other boards, pp. 63-64). In general,

the study succeeded in finding differences between boards' standards,[2] in determining how consistent the differences were across grade levels, and in determining the reliability of the results in the physics and mathematics studies. These results suggest that, for these two tests at least, cross-moderation procedures are potentially useful. The results of the study of the French test, on the other hand, were too inconsistent to allow any definite conclusions to be drawn. The reasons for these inconsistencies were probably the nature of the exams (i.e., the more subjective nature of judgments in foreign language exams than in subjects such as physics and mathematics), the different weights that each board applied to different components, and the extent of missing evidence (for example, the oral component of the tests). The effectiveness of cross-moderation in foreign language tests might improve if, for the oral component, specific elements a response should contain and standards for evaluation were articulated.

Based on the results of the study, the authors made several recommendations (pp. 68-71). First, they suggested that cross-moderation procedures continue to be used, but for purposes of identification rather than ratification. The modest improvement in reliability of grades given after marking over immediate-impression grades led them to conclude that the amount of time required for assessors to re-mark scripts is probably not justified. However, they stressed the importance of assessors' having a thorough knowledge of the relevant boards' marking schemes. Further recommendations were that 4 boards be involved in each study, and that each assessor grade 40-50 scripts from each board in a 3-day session. Exploratory studies (such as Johnson and Cohen's) that result in reliability figures above 0.7 should be followed up with a more complete study that includes all boards. They recommended using a balanced incomplete block design to make the task more manageable. Finally, they concluded that

---

[2] Note that this finding of differences between boards' standards is unrelated to the finding mentioned above of no overall tendency on the part of assessors to be severe or lenient. In the former case, scripts taken from certain boards were graded more leniently or severely than scripts from other boards *by all assessors*, regardless of the board they were from. In the latter case, no assessors were found to give grades that were consistently lenient or severe to all scripts, regardless of the board they were taken from. These findings indicate that the boards' standards were different and that assessors were able to adapt to each board's level of leniency or severity.

> cross-moderation does always have the advantage over other methods that it provides a spin-off benefit in providing the people actually involved in the boards' assessment and, sometimes, grading procedures with an opportunity to meet together, to gain invaluable knowledge of other boards' schemes, and to argue amongst themselves the legitimacy of any differences they thereby discern in the different views of subject performance which they variously hold. (p. 62)

Bardell et al. (1978) recommended using scripts for which there is a high degree of agreement as benchmarks for use in training (p. 36). It may also be helpful to include outside subject specialists who could serve to help with specifying criteria and who may make the exercise more credible (p. 30).

Finally, Forrest and Shoesmith (1985) identify a number of questions that should be answered if a cross-moderation procedure is to be used (pp. 43-45). These questions include: (a) whether the aim of the study is ratification or identification; (b) how many boards are to be included; (c) what kind of sample is to be drawn; (d) whether there is a common trait that represents achievement in the subject; (e) who the moderators should be; (f) what the moderators should be asked to do; (g) whose criteria the moderators should use in making their judgments; and (h) how the results of the studies are to be reported, interpreted and applied.

Moderation by inspection would seem to hold promise for assessment systems under consideration in the U.S. that would rely on judgmental scoring of student work at the local school level. Comparison of samples of scores assigned by local teachers to those assigned by district or state level moderators would probably enhance the credibility of the locally assigned scores, provide a means of communicating common standards of performance, and increase the comparability of results across schools. The experience with this approach in England, however, makes it clear that systematic moderation by inspection requires substantial investment of time and resources. Since there are many variations in the specific procedures (e.g., heavy or light sampling of scripts, standards of agreement, and uses of discrepancies), it is also clear that considerable planning and communication would be required for successful implementation.

Cross-moderation compounds the problems of moderation by inspection within a single educational jurisdiction (e.g., district or state) by adding

differences between jurisdictions in such relevant areas as content coverage, assessment procedures, and scoring criteria. Nonetheless, if claims that state or district level results meet common national performance standards are to be made and taken seriously, then it seems likely that some means of judging the degree to which equally stringent standards are actually being employed across jurisdictions will be needed. Cross-moderation provides one potential model for attempting to achieve this end.

Statistical Moderation

Statistical moderation procedures involve the use of information from an external moderating instrument to adjust teacher-assigned grades on an internal examination. The moderating information usually consists of scores on a reference or monitor examination, but can also include other sources of information, such as age, gender, SES level, past grades, etc. (some studies that involve the use of these kinds of information are dealt with more fully in the following section, "Other Moderation Procedures"). The rationale behind statistical moderation procedures is that the teacher-assessed components of the examination are likely to be more valid in terms of ranking the students, while the external exam is more suitable for establishing the relative standard of work across schools (Cohen & Deale, 1977, p. 49). The major assumption made in statistical moderation is that "the teacher is assessing, over a period of time, essentially the same skills and abilities as are assessed in the external examination" (p. 49). However, Smith (1978) points out, "the degree of overlap or correspondence is a contentious point: too little and the moderating instrument is unsuitable, too much and doubt is cast on the advisability of having both components as part of the same examination process." Both Smith (p. 20) and Cohen and Deale (1977, p. 49) suggest that correlations below 0.50 or 0.60 are probably too low. In addition, because the average group performance is used to determine whether marks need to be adjusted, statistical moderation will only be effective if the groups are fairly large since, in a small group, one invalid grade can have a substantial effect on the correlation coefficient (p. 21).

Cohen and Deale (1977) identify two ways that statistical moderation can be used. In the first, it is assumed that "the average grade . . . of candidates from a particular school should be the same, within statistical limits, for both the teacher assessment and the external assessment; if they are not, adjustments

are made to the school grades or marks so that the average does come within the tolerance limits." However, they point out, this method assumes that there are no real differences between the performance on the internal and external components in any particular school. In the second way of using statistical moderation, it is the first step in a process that combines statistical moderation and moderation by inspection. If a school's average grades do not fall within the specified statistical tolerance limits, additional information is gathered in order to determine whether grade adjustment is necessary (p. 50). Procedures of the latter type will be discussed in the following section.

Smith (1978, pp. 22-23) identifies three criteria that an examination should satisfy in order to be used in moderation. First, it should be reliable and capable of being marked with a high degree of consistency. With regard to the latter criterion, an objective test would be more suitable as a moderating instrument than an essay test, for example. As a corollary to this requirement, the external component should make up as large a part of the examination as possible. Second, as mentioned above, the moderating test should measure, to a reasonable degree, the same skills and abilities as are measured by the teacher on the internal component of the examination. It should be noted that, if an objective test is used, the moderating instrument can meet the consistency criterion, but at the expense of similarity to the internal component. Third, the test should be as fair as possible to all candidates; that is, it should validly reflect the skills and abilities emphasized on the internal assessment. The extent to which an external examination can be confidently used, without recourse to additional information, depends in large part on the extent to which the exam satisfies the above criteria. Because no exam can ever fully satisfy them, Smith argues that it is "probably just as unwise to place all one's faith in the moderating instrument and to adjust candidates' internally assessed marks in strict accordance with performance in it as it would be to accept the internally assessed marks without applying any kind of moderating technique" (p. 26). Small differences between group performance on the internal and external components of the examination can be due to error in the moderating instrument or to differences in the interest and skills of the students. He therefore advocates a "midway position" in which action is taken only if grades on the internal and external components are sufficiently different, that is, if internally assessed grades fall outside established tolerance limits. The use of tolerance

limits is especially important in the case where small groups are involved, both because an atypical grade can have a disproportionate effect on the overall results, and because a teacher may not have enough information to compare reliably his or her students to an absolute standard (pp. 26-27).

If the method involving the use of tolerance limits is chosen, how are the tolerance limits to be calculated? There are a number of factors that will affect these limits (Smith, 1978, p. 27). First is the extent to which the external component of the exam satisfies the three criteria listed above. Other factors that should be considered when calculating tolerance limits are the correlation between the internally and externally assessed components, the spread of marks in the two sets (as indicated by the standard deviations) and the number of candidates in each assessment set. Because these three factors are likely to vary from school to school, it may not be advisable to use the same tolerance limits for all schools (Smith, 1978, p. 27; Cohen & Deale, 1977, p. 51). On the other hand, calculating limits separately for each school would "produce a bewildering array of adjustments" (Smith, 1978, p. 27).

Once a decision has been made to adjust grades, the method of adjustment must be chosen. Smith (1978, pp. 23-25) delineates two methods by which statistical moderation can be accomplished. The first, which he terms scaling, corresponds to linear equating: "The marks from the internal assessment for each assessment are . . . adjusted to give the same mean and standard deviation as the distribution of marks for the moderating instrument of the candidates in that group" (p. 23). The second method, mapping, corresponds to equipercentile equating:

> The results on the moderating instrument of all candidates from each centre or assessment group are ranked. The candidates are also ranked in the order determined by the internal assessment. The top candidate on the internal assessment is then given a mark equivalent to the top mark obtained in the group on the moderating instrument, the next highest moderating test mark is given to the candidate ranked second by the centre, and so on down the rank order for the internal assessment. (p. 24)

In each case, the teacher's rank ordering of the students is unchanged, but in comparison to students with different teachers their relative standing may increase or decrease because grades are adjusted to conform to the common

moderating instrument. Cohen and Deale (1977) identify a mapping strategy that is particularly simple to use, especially if the number of students is small, in which teachers submit only a rank ordering rather than a set of grades. "A candidate in any position in this order is then awarded a mark equal to that obtained on the external examination by the candidate who was in that position in the order determined by the results on the external examination." Where schools are asked to submit grades, these grades are accepted if they fall "within specified limits of the average obtained by the school's pupils in the appropriate part of the external examination" (p. 51).

Another question that arises, with respect to the use of tolerance limits, is, if grades need to be adjusted, whether they should be adjusted to just within the tolerance limits or so they are as close as possible to the average external examination grade (Cohen & Deale, 1977, p. 52; Smith, 1978, p. 34). According to Cohen and Deale, the latter is the preferred method since the former "seems to benefit the pupils of the lenient assessor and penalize those of the teacher whose assessments are severe" (p. 52). Smith (pp. 34-35) acknowledges this problem, and adds that a teacher could obtain unfairly high grades for his or her students by awarding very generous grades since the moderation procedure would not fully adjust for the discrepancy, but he also points out a problem with adjusting scores so they are as close as possible to the mean. If a center's grades fall just outside the tolerance limits, the grades for all its students will be lowered. However, a center whose grades fall just within the limits will not require adjustment. This procedure differentially punishes students from centers whose grades fall just outside the tolerance limits. Either way of using tolerance limits suffers from validity problems. Thus, it is hard to justify the use of tolerance limits rather than simply adjusting grades assigned by all teachers so that the means based on internal assessments equal those based on the external one. A final point in regard to grade adjustment concerns the tendency for marks awarded on the internal assessment to be higher than those on the external assessment (Smith, 1978, p. 36). In order to compensate for this tendency, the mean difference between the two sets of marks is calculated for each center, and the average of these mean differences is taken. This average is then deducted from the difference between the mean marks for each center and these corrected differences are used.

A number of additional questions and issues regarding statistical moderation are discussed. One question is what percentage of the overall examination the moderating instrument should comprise. Smith (1978, pp. 21, 22) suggests that the moderating instrument should make up as large a portion of the overall examination as the internally-assessed component in order to minimize measurement error in the moderating instrument. If the internally assessed component represents the whole of the exam, a suitable external criterion needs to be found for moderating.

The need to supply opportunities for teachers to get help and information about the moderation procedure is emphasized by several authors. Cohen and Deale (1977) suggest that training sessions be offered to teachers to enable them to meet and discuss assessment and moderation procedures with other teachers and with moderators (p. 50). Smith (1978) suggests that it may be advisable to organize geographically compact local centers. Information should be supplied to each center about how the moderation procedure has affected their internally-assessed grades to enable teachers "to build up a bank of experience which will assist them in future years in making their assessments" (p. 38). Finally, the authors of both reports stress the necessity of verifying decisions made about borderline cases, especially if moderation will result in a negative adjustment (Cohen & Deale, 1977, p. 52; Smith, 1978, p. 29) and Smith recommends careful attention to the results of moderation in very small centers (p. 28).

A final problem that Smith (1978) discusses is that of authenticating the work submitted by the candidates (pp. 42-44). Authentication is particularly problematic in cases where candidates are required to submit final projects or studies that may be completed partly or entirely outside the classroom. It will be difficult, in these cases, for the teachers to ensure that the work is the candidate's own or that the candidate has accurately reported any assistance received or sources used.

A survey of teachers conducted in 1975 was discussed above in the section on cross-moderation procedures. Smith (p. 38) gives results from the same survey about the attitudes of teachers toward statistical moderation procedures. In general, teachers found the moderation procedures to be acceptable. However, Smith suggests that their support derives from support for the assessment procedure and their perception that no other viable alternative moderation procedure is available.

Smith advises that potential moderation procedures be considered when an assessment procedure is being designed: "Although the assessment procedure is not tailored to meet the needs of a particular type of moderation procedure, it would be foolish and irresponsible not to bear in mind what moderation procedure(s) can be used in the context of the scheme under consideration" (p. 40). It will often be necessary, because of limitations in the resources that are available, to compromise between the ideal moderation procedure and the easiest and least expensive solution, namely, no moderation at all (p. 41).

Later studies dealing with statistical moderation express some disillusionment with the procedures. The consensus opinion appears to have shifted away from statistical moderation and towards elaborated methods of cross-moderation or moderation by inspection (see the following section). Bardell et al. (1978) identify an overwhelming array of forms that statistical moderation can take, decisions that must be made, and problems that can arise. First, they point out that

> the monitor test may take one of several forms. It may be a subject attainment test or it may be of a general kind, testing academic aptitude or reasoning skills. In the former case it may be an integral part of the normal examination or be set as a supplementary test. The monitor test may be set to a sample of candidates or to all candidates. Depending on its form it may be marked subjectively (for example in essay form) or objectively (for example in multiple-choice form). (p. 19)

As pointed out by Smith above, a monitor test can only work to the extent to which it is relevant (i.e., closely related to the exam being moderated) and fair (p. 20). The use of a relevant monitor test will have two effects (p. 21). First, it reduces the uncertainty surrounding the study since confidence in the results is a direct function of the correlation between the monitor and internal examinations. Second, it helps to increase the importance of the study since it is not of interest that the monitor predicts no difference in grades if the correlation indicates that the monitor fails to explain a large proportion of the variance in the grades.

In regard to the second criterion listed above, fairness, Bardell et al. point out that judging fairness is not easy to do. It is often difficult to distinguish differences that are due to bias and differences that reflect what the test is designed to measure. In addition, a given test may favor some candidates in

some ways and other candidates in other ways (p. 21). The use of an aptitude, rather than a subject-based test, would seem advisable in this respect. Performance on aptitude tests is less likely to be affected by syllabus differences than performance on subject-based tests; thus, aptitude tests would seem to be less biased than a subject-based test that was better aligned with the curriculum of some boards than with that of others. However, because aptitude tests are likely to be less relevant to the subject being assessed than a subject-based test would be, it would be more difficult to make valid inferences from the results of an aptitude test.

A potential technical problem with statistical moderation arises if the relationship of internal and external assessments varies from board to board. This indicates that differences in standards between boards are not constant across the grade range (i.e., boards may be more likely to be lenient or severe for more able candidates than for less able candidates) (p. 21). Other problems with the administration of monitor tests include the expense and time-commitment they require, the difficulty of finding adequate samples, and the necessity of obtaining the schools' cooperation and goodwill (p. 24). A possible alternative to a supplementary, external test is the use of a common, internal subject-based paper; this option would not require supplemental testing. However, even if all boards use a common paper, there is no guarantee that the standards applied are comparable (p. 25).

Studies of moderation procedures dating from the mid-1980's no longer deal in depth with statistical moderation. Discussions of statistical moderation are brief and, in general, serve only to explain why such procedures have not been found to be satisfactory for the purpose of moderation. Johnson and Cohen, for example, point out that the use of the reference test method is based on two assumptions: first, that performance on the reference test will be strongly related to performance on the exam under review, and, second, that the reference test does not favor any group of students over the others (1983, pp. 3-4). They point out that: "It was doubt about [the validity of these assumptions] which led, in 1976, to the discontinuation of this particular approach to grade comparability investigation" (p. 4). Similarly, Forrest and Shoesmith (1985, p. 11) point out that while the use of reference tests may show that the candidates from one board are more able than the candidates from another board, it won't tell how large the difference is. In addition, they point out the

extreme difficulty and expense involved in finding tests that are sufficiently fair, reliable and relevant to perform adequately as reference tests.

The shortcomings of statistical moderation procedures that have led to a disenchantment with the approach in England are likely to be considered even more serious in possible applications in the U.S. Reliance on an external examination to make adjustments in the results of internal assessments is likely to undermine the goals of the internal assessment and distort instruction. An external or monitor assessment may be valuable if used to identify situations where additional information or the use of audits or inspection moderation procedures may be needed. Such hybrid approaches are discussed in the following section.

Other Moderation Procedures

Bardell et al. (1978) summarize the problems involved in trying to moderate results across examination boards (pp. 15-18). Their discussion focuses on trying to moderate exam results without recourse to any other information. First, they point out that it can't be assumed that the distributions of grades for each board should be the same: Some boards are certain to have candidates that are of higher ability than those of other boards. While it is possible to adjust statistically for between-board differences in grade distributions, the extent to which judgments made on the basis of these analyses are valid depends on two assumptions: first, that exams in a given subject from different boards are measuring the same thing and, second, that the candidates who take a subject exam from one board are of no greater or lesser ability than those who take exams in the same subject from other boards. One way to assess the extent to which the second assumption is met is to do an analysis using candidates who take the same exam in more than one board. However, these candidates are not typical of all candidates; they are more likely to be on the border between pass and fail. An additional problem is that any grading differences found in studies of this kind can be due either to differences in standards *or* to differences in the ranking of the students. If students are ranked differently by the results of the two exams (as seems almost certain), it is extremely difficult to determine to what extent the standards of the two boards are different.

Another way to explore comparability between boards is to compare the grade distributions in two different subjects for the same group of examinees. It

is not sufficiently clear, however, that the two grade distributions *should* be the same. One case in which it might be possible to reach a justifiable conclusion would be if, for all those taking exams in both English and history, the average performance in history was a half a grade higher than the average performance in English in all but one of the boards. One might then conclude that the last board's standards are out of line in English, history, or both. Even in such an extreme case, however, it would still be possible that the finding was due to the deviant board's placing different emphases on the subjects than the other boards. Moreover, most situations are quite unlikely to be this clear cut. The problems outlined above demonstrate the need for some type of additional, possibly external, information to make moderation possible.

The studies that remain to be discussed all describe moderation procedures that cannot be classified into any of the categories described in the previous sections: moderation by inspection, cross moderation, or statistical moderation. The procedures discussed are either extensions of one of these types of moderation or combinations of them. Nuttall & Armitage (1985) discuss a variety of issues that arise in developing a moderation procedure and make a number of suggestions based on their findings. The discussion is based on the experiences of the Business and Technician Education Council (BTEC), and some of the definitions of certain methods differ somewhat from those described in previous sections. For example, the authors distinguish between moderating instruments, which consist of an external examination, and "statistical monitoring," in which standards are adjusted on the basis of "internal" information about the students obtained from TEC[3] records, such as performance on other TEC units (p. 9).

The authors suggest that a moderating instrument should be used only for identification of suspect cases for which follow-up verification will be undertaken, not as the arbiter of standards, since there are many factors that may influence how a student does and no moderating instrument can possibly detect all of them (pp. 3, 10). They also suggest that a definition of standards of performance should not be based solely on performance or output but should also include "the design of the course, its relevance to the needs of the 'consumer',

[3] Technician Education Council. In 1983, TEC and BEC (Business Education Council) combined into BTEC; the study was commissioned by TEC and makes use of TEC data but was not completed until after the formation of BTEC.

[and] the physical and human resources supporting it" (p. 4). Because the endorsement of students' grades as agreeing with national standards was viewed as the most important aspect of standards by the moderators (p. 5), it was the main focus of the study described below. Even with this narrowed focus, however, the study found that "the easily used phrase 'national standards' is, in practice, difficult to define and interpret. A research project on possible deviations from a national standard has to employ an operational definition of that standard" (p. 8). This problem was avoided, but not solved, by using the national average as the standard. Thus, for example, if, nationally, 43% of candidates in a particular unit received a Pass with Merit or better, 43% Merit or better was used as the standard for that unit and centers' percentages of Merit or better were compared to that figure.

In developing a moderating instrument, two decisions need to be made about the nature of the examination. First, the test can be separate from the within-school assessments or its results can be used in awarding grades (p. 9). If exam results are not used in awarding grades, it is difficult to get students to take the exam seriously. On the other hand, use of the results in grading may cause teachers to teach to the test, resulting in distortion of the program content (p. 19). Second, a decision has to be made as to whether the test should be similar to within-school tests and specific to the content of the course or more broad.

For purposes of the study, two versions of the moderating instrument were developed, one with a broad test and the other with a specific test (p. 11). Each version also included performance on TEC units already taken and the age of the student. Other sources of information (for example, gender and the occupational relevance of the students' studies) were not found to be significant predictors (p. 13). The study's aim was to predict the percentage of students in a given class who would achieve a Merit or better and compare it to the actual percentage (p. 14). The criteria for determining how much discrepancy would be allowed before standards were to be flagged as potentially deviant were somewhat arbitrary (p. 16). The setting of these criteria would, of course, have a considerable effect on the results of the exercise.

Nuttall and Armitage describe the basis on which the usefulness of the moderating instrument was to be judged:

> It is clearly a vital test of the feasibility of a moderating instrument to establish whether, when the predicted and actual number of Merit or better grades differed substantially, grading standards were at fault or whether there was another plausible reason or set of reasons that could explain the result. The moderating instrument is hardly likely to be infallible but, if it is to be of any value, a reasonable proportion of the deviant cases it highlights must be cases where those involved accept that their grading standards are somewhat out of line, and must not be capable of being 'explained away.' (p. 15)

The results obtained from the instrument were validated by asking a panel of full time moderators to visit a number of the schools and come to an independent conclusion about whether grading standards were in line or not. The important test would be whether the deviant findings were consistent between the moderators and the moderating instrument (p. 16).

The results of the study were as follows: Using the specific test, 9 out of 69 classes (13.0%) were labeled as deviant; using the broad test, 19 out of 182 (10.4%) were found to be deviant (p. 16). Because the study design tried to take the ability and motivation of the students into account, a convincing explanation for rejecting a finding of deviant grading standards would need to be based on circumstances specific to a certain class (p. 17). An example of such an explanation might be if one class took the test used for moderation earlier than other classes, prior to being exposed to some of the relevant information (p. 15).

College staff and moderators were given the opportunity to comment about deviant findings and their comments were classified according to whether or not they confirmed the results of the study:

	Positive confirmation	Not proven	Rejected	Total deviant cases
Broad test	6 (31.6%)	4 (21.1%)	9 (47.4%)	19
Specific test	5 (55.6%)	1 (11.1%)	3 (33.3%)	9

The pattern was similar among the subsample of colleges that were visited by the moderators (25 out of the 130 colleges that participated). In addition, the moderators did not detect any deviations among those 25 colleges that the

instrument failed to detect (p. 18). The possibility of correctly identifying deviant cases is one potential advantage of using a moderating instrument. However, it also has utility for deterring schools from deviating from national standards and reassuring the public that standards are being monitored (p. 18).

Nuttall and Armitage conclude by discussing a number of issues that arise in connection with the use of a moderating instrument. First, they revisit the question of whether the test should be specific to an area or more broad. They find that the specific test, either alone or combined with age information, is somewhat more efficient than the broad test, on its own or combined with age (p. 22). The major disadvantage of using an external area-specific test is, of course, the difficulty and expense of developing the tests (p. 22). On a somewhat more positive note, they argue that, if the program is implemented, "not proven" verdicts will become less of a problem over time as evidence of deviance or of alternative explanations is gathered at those schools (p. 19). In addition, they claim, feedback to schools would improve over time. More detailed information would become available and explanations for deviant findings would become more refined and plausible (pp. 19-20).

The authors emphasize that, in order to allay public concerns over educational standards, it is necessary to have some sort of monitoring system that both the public and professionals will have confidence in. However, the monitoring system should not be implemented at the expense of local control. Because compulsory use of a wholly external test may result in bias and distortion of teaching practices, they did not recommend this option (p. 20).[4] One possible way to achieve an acceptable degree of comparability while still maintaining flexibility may be to develop a question bank from which schools can choose sets of questions (p. 21). In any case, they believe, the provision of more information about programs, assessment and grades could serve only to enhance public respect for the system. In addition, in order for a moderation system to be effective, it is crucial that findings of deviation be followed up by an informed group of moderators (p. 23). The study verified that the use of both an external examination and internally-obtained information was to be preferred over the use of either option alone (p. 22). As a final caution, however, Nuttall and

[4] Since the time of this study, BTEC has changed its moderating instrument. The instrument is now in the form of a task or assignment that is carried out as an integral part of the course. The use of the task in assigning grades is optional (Nuttall & Thomas, 1992, p. 9).

Armitage point out that "a moderating instrument will never be infallible, and to be effective it must supplement and complement other systems of monitoring and moderation. The price of reconciling local needs with national standards will never be cheap" (p. 24).

In 1992, the Business and Technology Education Council undertook a study of how well their moderating procedure was working (BTEC, 1992). Each year, samples of students from targeted core modules take part in a case study which is graded by independent assessors. The grades on these case studies in combination with data on age and performance on the first year core module are used to predict grades awarded on the final year core modules. Centers for which there are substantial discrepancies between predicted and actual grades are visited by advisers from the Business and Technology Education Council, who try to determine the cause of the discrepancies (p. 1).

Several possible explanations for the discrepancies are identified: (a) the quality of the program is higher or lower than average; (b) the prediction may not reflect all circumstances and factors; or (c) grading standards in the center are not in line with national standards (pp. 1-2). The results of the 1991 case study indicated that 8.3% of classes were not within the range of predicted grades. Of this percentage, 59% were due to discrepant grading standards, 26% were probably due to the quality of the program, and 15% were the result of prediction errors or were unexplained (pp. 3-5).

One of the causes of grading inconsistencies that was identified was that some internal staff involved in assigning grades were not fully integrated into the program team, resulting in a lack of internal consistency. This problem could arise any time there were part-time or new staff, changes in staff, or a single specialist in a given area. In addition, lack of moderation procedures within the schools was given as a possible cause. Other possibilities had to do with differences in how the assessments were used in grading. In some cases, a single assessment was used for a large proportion of the grade. The balance among time-constrained tests, projects, and assignments varied from one program to another. Finally, assessment criteria were not consistent across programs.

Centers' use of grading standards that were out of line was a final source of inconsistency discussed. Three main factors in relation to inconsistency in

grading standards were identified. First, expectations of the students may have been set too high or too low. Second, in many cases, there was no reference point for grading standards identified. Finally, the relationship between staff and students or student attitudes may have influenced grading (pp. 5-8).

As a result of these findings, the report concluded that the existence of centers whose grading standards are out of line attests to the need for continued monitoring of centers' assessments. It was recommended that centers take into consideration a number of suggestions. These suggestions include (pp. 11-12):

- Ensure that all centers' program teams have a clearly defined structure with individuals identified as assessors and coordinators.
- In the case of newly formed teams provide guidance and assistance on assessment and grading procedures.
- Check that there are effective procedures to monitor inexperienced or part time staff.
- Establish additional internal moderation arrangements if staff take over a module mid way through.
- Identify any lone specialist within the team and ensure there are arrangements for comparison of grades.
- Monitor and Review internal moderation procedures. Ensure these involve all team members and are effective in identifying inconsistencies within the team.
- Develop and maintain benchmark examples of graded work to assist new staff in interpreting standards.
- Develop assessment criteria which define the performance expected for each score and enable grading decisions to be open to scrutiny.
- In collaborative projects build in formative assessment and a contribution profile to ensure that grades are based on individual students' work.

Nuttall and Thomas (1992) discuss a moderation study done for the National Vocational Qualifications and Scottish Vocational Qualifications. The report was designed to address the issue of comparability in the standards applied by different examination centers. The authors discuss a method of statistical process control and how it can apply to the question of comparability. Statistical process control "encourages the study of fluctuations from the desired product quality, recognizing that much fluctuation is attributable to random

causes and requires no corrective action" (p. 3). The authors elaborate several strategies by which misinterpretation of standards can be minimized (pp. 4-5). One way is through the training of assessors and verifiers, who would be required to be certified. A second way is to verify that assessors and internal verification and administrative procedures meet quality criteria. A third would be to employ external verifiers who are responsible for assessment in a number of different centers. And a final strategy would be to implement a monitoring procedure to be used nationally. In the study, the authors explore one such procedure, which they term a monitoring procedure based on performance variables.

The monitoring procedure is based on the following premises (p. 11): (a) It is designed to enhance quality and national credibility of standards and qualification; (b) it focuses on uniformity of the interpretation of standards; (c) it is used as a screening device, not as an arbiter; (d) it applies to a given sample of candidates for a given award on any one occasion; and (e) it applies to samples of qualifications and needs to be tailored to different contexts (i.e., different levels and occupational sectors). According to this approach, a judgment of competence is a function of (pp. 11-12): (a) a measure of the candidate's competence; (b) the candidate's prior attainments; (c) the quality of the experiences leading to the acquisition of competence; (d) other characteristics of the candidate (age, sex, etc.); (e) the candidate's motivation to do well; (f) relevant characteristics of the center; and (g) misinterpretation (if any) of the standards of competence. Many of these factors are very difficult to assess and, therefore, evidence should be carefully and thoroughly reviewed before a judgment of misinterpretation of standards is made. In addition, inconsistent assessors are difficult to detect; this possibility should be carefully considered as well (p. 14).

In developing a monitoring procedure for the vocational qualifying exams, the authors identified a couple of important issues that needed to be considered (pp. 15-16). First, what should be used as the outcome variable? For the vocational qualification exams, examinees are categorized simply as competent/not yet competent. A judgment of not yet competent encompasses a large array of possibilities, so the use of this outcome variable involves a number of theoretical and statistical problems. Second, should the monitoring procedure include a centrally prescribed assessment? The advantages of such an assessment include that it improves the ability of the monitoring procedure to

assess competence and that it increases the credibility of the exercise. The main disadvantage is the expense involved in creating the examination.

If a centrally prescribed assessment is to be used, it must meet the following requirements (p. 18): (a) It must be a task or activity that can be assessed in a uniform fashion across centers; (b) it must be a task or activity that can be administered in a reasonably standardized fashion across centers; and (c) it must be a task or activity that is a valid assessment of the target units and/or elements. To develop the assessment, the units or elements to be used must be identified; these units or elements must be ones that (p. 19): (a) are likely to be attained towards the end of the typical sequence of unit accumulation; (b) are regarded as key or particularly important facets of the total competence; and (c) are amenable to assessment meeting the criteria above.

Nuttall and Thomas discuss some other options that could be used instead of a centrally prescribed assessment that has been developed specifically for use as a monitoring procedure (pp. 20-21). They point out that the use of national reference tests (aptitude tests, for example) suffers from a lack of face validity. However, their use would be appropriate in areas in which core skills play a major part. The advantage of such tests is that they could be much more widely utilized and, therefore, would be cheaper. Another option identified (p. 21) is to use the centrally prescribed instrument as the criterion measure, rather than as one of the independent variables. In the regular monitoring procedure model, the assessment score, along with other available information, is used to predict a center's level of performance and this predicted performance is compared with actual performance. The other option would be to determine whether the local judgment of competence was in accord with the judgment of competence on the centrally prescribed assessment and, if not, to conduct a follow-up investigation.

In discussing variables that could be used as part of a monitoring procedure, Nuttall and Thomas distinguish between predictor variables, which predict performance on the outcome variable, and explanatory variables, which are associated with a center's misinterpretation of standards. They describe the criteria against which to judge potential variables (pp. 22-23): (a) They should be available or readily collectable; (b) they should be measurable and comparable; (c) they should be reliable and valid; (d) they should be demonstrably related to the outcome measure; and (e) they should be politically acceptable. The list of variables that they believe may prove useful in predicting

outcomes on the vocational qualifying exams (pp. 23-29) includes candidate measures (such as prior achievement, percentage of time allocated for training, percentage of time absent from training/work, candidate's assessment of training experience, age at award, sex, ethnicity, etc.); center variables (such as geographical region, center-based assessment resources, timing and scheduling of assessment, technique of assessment, percentage of unit assessments successful, etc.); aggregated assessor and internal verifier characteristics (such as mean years of experience in vocation/occupation, experience as an assessor/verifier, assessor/verifier turnover, etc.); and aggregated candidate characteristics (such as percentage of ethnicity, percentage ESL, etc.). It should be noted that the use of a number of these variables (such as sex, ethnicity, center-based resources, and percentage ESL) would not be politically feasible in the U.S. if their use resulted in implicitly accepting different standards for different groups.

Once data have been gathered on the variables selected, a statistical method must be chosen that can help identify centers that may be misinterpreting national standards. Simply comparing each center's results against the national average is unlikely to be useful for two reasons. First, some centers may be too small to give accurate aggregated results, and second, this method cannot take into account relevant background factors that may explain any discrepancies (p. 29). Instead, the authors recommend the use of multilevel modeling which is a sophisticated form of regression analysis that allows data from more than one level (for example, candidate level data and center level data) to be taken into account simultaneously (pp. 30-31). The use of multilevel modeling will allow significant predictor variables and explanatory variables that are associated with the expected outcomes to be identified. These variables would then be used in the screening instrument against which centers' interpretation of standards would be assessed (p. 33).

James and Conner (1992) describe how moderation helped to establish consistency of assessment procedures within and across schools and between Local Education Authorities (LEAs). The evidence given is primarily based on observations done at times convenient to the teachers and researchers and does not give a comprehensive or representative description of what was happening in the four LEAs. The results, therefore, represent reflections on some general issues that arose. The authors also discuss other methodological issues (validity,

for example) as well as practical and professional issues related to moderation (p. 5).

The researchers' observations of moderating procedures at a number of schools gave rise to the following major issues:

1. Concerns were raised about the validity of both standardized test results and teacher assessments. The authors argue that, if the problem concerns the validity of standardized tests for measuring what they are supposed to measure, the test developers must address the problem (p. 20). However, in terms of teacher assessments, there is more room for help from moderators. Unfortunately, the scope of the task and time demands on moderators make it difficult for moderators to give guidance to teachers. This is one area in which additional research and increased professional development and training would be beneficial (pp. 20-21).

2. Reliability in the form of consistency in teachers' presentation of tasks was a problem. While greater standardization of assessment tasks would help, it would also decrease the validity of the tasks by increasing their artificiality (p. 21).

3. Reliability in terms of the consistency in teachers' interpretation of standards was difficult to assess because of the small scale of the project. However, where comparisons were possible, the interpretations and judgments made were similar and common problems and concerns arose (p. 21).

4. A major problem concerned the teachers' success in applying a criterion-referenced rather than a norm-referenced model of assessment. "Since teachers' perceptions of what is 'normal' is conditioned by their particular experiences, which are rarely representative of all schools, such **normative expectations** are a threat to consistency in interpretation and therefore the reliability of the assessment results" (p. 22, emphasis in original).

5. The difficulty the moderators had in interpreting and balancing the various expectations placed on them was a problem. Decisions had to be made about what to concentrate on which increased problems of inconsistency. Further clarification and simplification of moderators' roles is necessary to solve these problems. However, it is possible that limiting the role of moderators may also serve to reduce the satisfaction of doing the job (pp. 22-23).

6. Another question that arose concerned who the moderators should be. Pre-existing roles and commitments of moderators had an impact, sometimes positive and sometimes negative, on how effectively they did their job. However, the quality of individual moderators is a more important issue. The findings of the study suggest that the single most important attribute a moderator must have is credibility (pp. 23-34).

7. The training of moderators led to inconsistencies in how they did their jobs. Because different moderators had different backgrounds and qualifications, they required different emphases in their training (p. 24). In addition, there was a tension between specificity and comprehensiveness in the training moderators received.

8. Finally, timing was a problem since moderators had more time available for training before the standardized testing occurred. This was not the ideal time for training to occur since many problems arose in administration of the tests that were not anticipated in the training (p. 25).

The study ends by discussing some general recommendations to ensure the smooth running of the moderation procedure. First, the authors stress the importance of informal interaction between teachers for increasing consistency. They recommend that opportunities be provided for teachers to meet and discuss assessment issues. However, they note, this will not suffice for ensuring comparability so some formal structure is still necessary. As an aspect of this structure, the system needs to encourage communication among all the participants, including between teachers and standardized test developers (pp. 26-27).

Another recommendation concerns the development of a system of accreditation to ensure consistency across schools. This accreditation should include endorsement of the accredited schools' assessments. The best way to encourage in-school development and consistency across schools is to develop a formal structure or organization for monitoring schools. Such a system must be implemented at the level of the school or group of schools and might consist of an in-school moderator or linked sequence of moderators who remain in contact with moderators at other schools. It may also consist of a validation board for a school or group of schools which would ensure comparability of teachers' judgments and provide feedback, encouraging professional development in the area (pp. 28-29).

Conclusion

The previously mentioned observation of Nuttall and Armitage (1992) that some sort of monitoring system that both the public and professionals have confidence in will be necessary in order to make the claim credible that local assessments meet national standards seems just as applicable in the U.S. as in England. Neither a pure moderation by inspection nor a strict statistical

moderation system is likely to meet this need, however. Relying solely on moderation by inspection would probably have credibility problems in a country like the U.S where technological and statistical solutions are more the norm. Furthermore, this approach is apt to prove unwieldy in a country as large as the U.S. Strict statistical moderation, on the other hand, would undermine the goals of the assessment systems that are currently under consideration. Thus, it seems more likely that some sort of hybrid system will be required that relies on a combination of an external assessment and statistical comparisons to identify places where more detailed information, the use of audits, or the use of moderation by inspection is needed.

The issues that led to the development of the English examination system are similar to issues being discussed currently in the U.S. In both cases, changes in the exam system were desired in order to enable more local control over test content as well as to provide tests that reflect more accurately the type of work being done by students in school. While the goals themselves are worth pursuing, the use of such tests does raise some problems as well. The purpose of this paper has been to discuss the various issues surrounding one of these problems, namely, the need to find a way to enable comparison of results across schools and regions and to ensure public belief in the validity of such assessments. The foregoing summary of experiences in England introduces a number of important issues and questions that will need to be addressed in order to resolve the problem of linking examination results in this country.

References

Bardell, G.S., Forrest, G.M. & Shoesmith, D.J. (1978). *Comparability in GCE: A review of the Boards' studies, 1964-1977.*

Cohen, L., & Deale, R.N. (1977). *Assessment by teachers in examinations at 16+.*

Forrest, G.M., & Shoesmith, D.J. (1985). *A second review of GCE Comparability Studies.*

Goff, P. (1992). *1991-92 Moderating instrument: A report.*

James, M., & Connor, C. (1992). *Moderation at Key Stage One across four LEAs 1992.*

Johnson, S., & Cohen, L. (1983). *Investigating grade comparability through cross-moderation.*

Linn, R.L. (1993). Linking distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Learning Research and Development Center, and the National Center for Education and the Economy. (1992). *The New Standards Project, 1992-1995: A proposal.* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

National Council on Education Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people.* Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1991). *The National Education Goals report, 1991: Building a nation of learners.* Washington, DC: Author.

National Vocational Qualifications (1991?). *National standards for assessment and verification.*

Nuttall, D.L., & Thomas, S. (1992). *Statistical quality assurance for NVQs and SVQs: Monitoring procedures based on performance variables.*

Nuttall, D.L., & Armitage, P. (1985). *Moderating Instrument Research Project: A summary report.*

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston, Kluwer Academic Publishers.

Schools Council (1965). *The Certificate of Secondary Education: School-based examinations* (Examination Bulletin No. 5). London: HMSO.

Shepard, L.A. (1991). Psychometricians' beliefs about learning influence testing. *Educational Researcher, 20*(7), 2-16.

Smith, G.A. (1978). *JMB experience of the moderation of internal assessments.*

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.